Egyptian Journal of Forensic Sciences (2014) 4, 83–89

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Egyptian Journal of Forensic Sciences

journal homepage: http://www.journals.elsevier.com/egyptian-journal-of-forensic-sciences

ORIGINAL ARTICLE

Performance of some supervised and unsupervised multivariate techniques for grouping authentic and unauthentic Viagra and Cialis

Michel J. Anzanello [a,*], Rafael S. Ortiz [b], Renata Limberger [c], Kristiane Mariotti [c]

[a] *Department of Industrial Engineering, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99 – 5L andar, Porto Alegre, RS, Brazil*

[b] *Rio Grande do Sul Technical and Scientifical Division, Brazilian Federal Police, Avenida Ipiranga 1365, 90160-093 Porto Alegre, RS, Brazil*

[c] *Department of Pharmacy, Universidade Federal do Rio Grande do Sul, Av. Ipiranga, 2752, 90610-000 Porto Alegre, RS, Brazil*

Received 3 October 2013; accepted 25 March 2014
Available online 5 May 2014

KEYWORDS

Counterfeit medicines;
PCA;
Supervised Techniques;
Unsupervised Techniques;
FTIR-ATR

Abstract A typical application of multivariate techniques in forensic analysis consists of discriminating between authentic and unauthentic samples of seized drugs, in addition to finding similar properties in the unauthentic samples. In this paper, the performance of several methods belonging to two different classes of multivariate techniques–supervised and unsupervised techniques–were compared. The supervised techniques (ST) are the *k*-Nearest Neighbor (KNN), Support Vector Machine (SVM), Probabilistic Neural Networks (PNN) and Linear Discriminant Analysis (LDA); the unsupervised techniques are the *k*-Means CA and the Fuzzy C-Means (FCM). The methods are applied to Infrared Spectroscopy by Fourier Transform (FTIR) from authentic and unauthentic Cialis and Viagra. The FTIR data are also transformed by Principal Components Analysis (PCA) and kernel functions aimed at improving the grouping performance. ST proved to be a more reasonable choice when the analysis is conducted on the original data, while the UT led to better results when applied to transformed data.

1. Introduction

The commerce of counterfeit Phosphodiesterase type 5 (PDE-5) inhibitors for the treatment of erectile dysfunction presented a large increase in the last decade. From January 2007 to September 2010, 80% of the reports issued by the Brazilian Federal Police (PF) were associated with seizures of unauthentic Cialis and Viagra samples.[1] Sildenafil (Viagra®, Pfizer), tadalafil (Cialis®, Eli Lilly) and vardenafil (Levitra®, Bayer) are responsible for a significant portion of counterfeit seizures due to their high commercial cost and embarrassment associated with the underlying pathology.[2] Counterfeit PDE-5 inhibitors represent serious risks to public health since there is no certainty about active

* Corresponding author. Tel.: +55 51 3308 4423.
E-mail addresses: michel.anzanello@gmail.com (M. J. Anzanello), rafaelortiz.rso@dpf.gov.br (R. S. Ortiz), renata@ufrgs.br (R. Limberger).

Peer review under responsibility of The International Association of Law and Forensic Sciences (IALFS).

http://dx.doi.org/10.1016/j.ejfs.2014.03.004

pharmacological ingredients, pharmaceutical dosage forms, and origin of raw materials.

Several analytical techniques enable the identification of tadalafil and sildenafil in pure or pharmaceutical forms, including the physical control of tablets,[3] inorganic profile by X-ray fluorescence spectrometry (XRF),[1] organic pro-file by electrospray ionization mass spectrometry (ESI-MS),[4] and Infrared Spectroscopy by Fourier Transform (FTIR).[5–7] Such data have been successfully analyzed by means of simple yet efficient multivariate techniques, as Principal Component Analysis (PCA),[8–10] Cluster Analysis (CA),[11] and more recently Data Mining (DM) techniques.[12–14] The scope of the multivariate tools in forensic applications is typically to discriminate between authentic and unauthentic samples of seized drugs, in addition to finding similar properties in the unauthentic samples. In light that CA and DM are multivariate techniques that rely on different theoretical fundamentals, the aim of this study is to assess the performance of such techniques on data from analytical techniques.

In this paper, the performance of two groups of multivariate techniques frequently used for analyzing sample properties and inserting samples into authentic and unauthentic categories–supervised and unsupervised techniques–are compared. Methods associated with Supervised Techniques (ST) are applied on two groups of variables: independent, e.g., variables arising from analytical techniques; and dependent variables, e.g., labels of authentic or unauthentic samples. STs establish a relationship between independent and dependent variables, yielding a model to classify new samples into categories. DM methods are inserted in this category, and include *k*-Nearest Neighbor (KNN), Support Vector Machine (SVM), Probabilistic Neural Networks (PNN) and Linear Discriminant Analysis (LDA), among others. On the other hand, Unsupervised Techniques (UTs) do not require a dependent variable for modeling; instead, UTs search for patterns among the independent variables, and groups of samples are formed based on the structure of the variables. UTs include clustering techniques, such as the *k*-Means CA and the Fuzzy C-Means (FCM). In these propositions, the performance was tested of four STs (KNN, SVM, PNN and LDA), and two UTs (*k*-Means CA and FCM); such techniques are applied to Infrared Spectroscopy by Fourier Transform (FTIR) from authentic and unauthentic Cialis and Viagra, and the resulting classification accuracies are assessed.

The main contribution of this paper is to provide the researcher with a better understanding of some multivariate techniques typically used in Forensic and Biomedical applications. In addition, it is extected that this research will unveil some advantages and disadvantages of each technique to help the researcher in choosing the most appropriate technique for each nature of analysis. Finally, the use of the Silhouette Index–a well known metric for measuring clustering quality in multivariate analysis–is seen as a relevant contribution to the forensic science field.

2. Materials

2.1. Samples

Twenty-five samples of authentic Viagra® and 28 samples of authentic Cialis® were analyzed. Six authentic Viagra® tablets containing 50 mg of Sildenafil (SLD) were supplied by Pfizer Ltda Laboratories, and 8 authentic Cialis® tablets containing 20 mg of Tadalafil (TAD) were supplied by Eli Lilly to Brazil Ltda Laboratories. Twenty authentic Cialis® tablets (TAD, 20 mg) from 8 distinct batches and 19 authentic Viagra® tablets (SLD, 50 mg) from 6 distinct batches were purchased in local pharmacies (Dimed S/A Distribuidora de Medicamentos, Porto Alegre, RS, Brazil). In addition, 104 counterfeit samples were sent for forensic analysis at the PF (Porto Alegre, Brazil). All samples were analyzed by ATR-FTIR.

2.2. ATR-FTIR analyses

All experiments employed a Nicolet 380 FTIR Spectrometer (Nicolet Instrument Co., Madison, Wisconsin State, USA) equipped with DTGS (Deuterated Triglycine Sulfate) detector and Smart Orbit single reflection diamond. An ATR sampling device was employed in all experiments. The spectra from a small amount of sample positioned on the ATR crystal were measured, and the transmittance values were converted to absorption. Genuine and counterfeit tablets were prepared the same way: the tablets were crushed in a porcelain mortar, and the powder was tested in the ATR-FTIR device. No sample treatment was necessary for measurement. Some of the authentic and counterfeit Viagra presented a film coating whose fragments were removed from the sample after crushing. As for samples presenting no film coating, the coating became part of the sample in the form of homogenized powder.

Next, a sample portion was directly placed in the ATR element, and the same pressure was used for all measurements. Each mixture was sampled three times (triplicate). Each spectrum comprises 16 co-added scans measured at a spectral resolution of 4 cm^{-1} in the 4000–525 cm^{-1} range. Spectral data were acquired with EZ OMNIC software, version 7.2a (Nicolet Instrument Co.). After the measurement, the crystal was cleaned with acetone. An hourly background spectrum was obtained against air with a clean and dry ATR element, using the same instrumental conditions as the samples. No spectrum pretreatments were employed.

Representative spectra of SLD (authentic Viagra active pharmaceutical ingredient-API), in purple, and TAD (authentic Cialis API), in red, are depicted in Figure 1. Representative spectra of genuine and counterfeit Viagra and Cialis samples are presented in Figure 2. The most important peaks for TAD can be associated with C—O bonds in the 1700 cm^{-1} band and C—C bonds from the ketone group in the 1280 and 1172 cm^{-1} band. As for the SLD, the 1676 cm^{-1} peak can be correlated to C—N stretching (1690–1640 cm^{-1}); N—H bending appears at 1647 cm^{-1} and the 1490 cm^{-1} band is the result of C—C bonds in a ring; C—N bonds in the O—C—N functional group absorb at 1400 cm^{-1} which accounts for the 1402 cm^{-1} absorbance; and the aryl C—N bonds are responsible for the 1269 cm^{-1} peak. In addition, there are some characteristic infrared absorption peaks for lactose (excipient of authentic Cialis) in 1048 cm^{-1}, 909 cm^{-1} and 890 cm^{-1}. The spectra for Cialis (TAD) and Viagra (SLD) consist of 300 and 177 samples (both in triplicate), respectively, and 661 variables (wave number).

2.3. Multivariate techniques

The following presents the fundamentals on the supervised and unsupervised techniques evaluated in this paper.

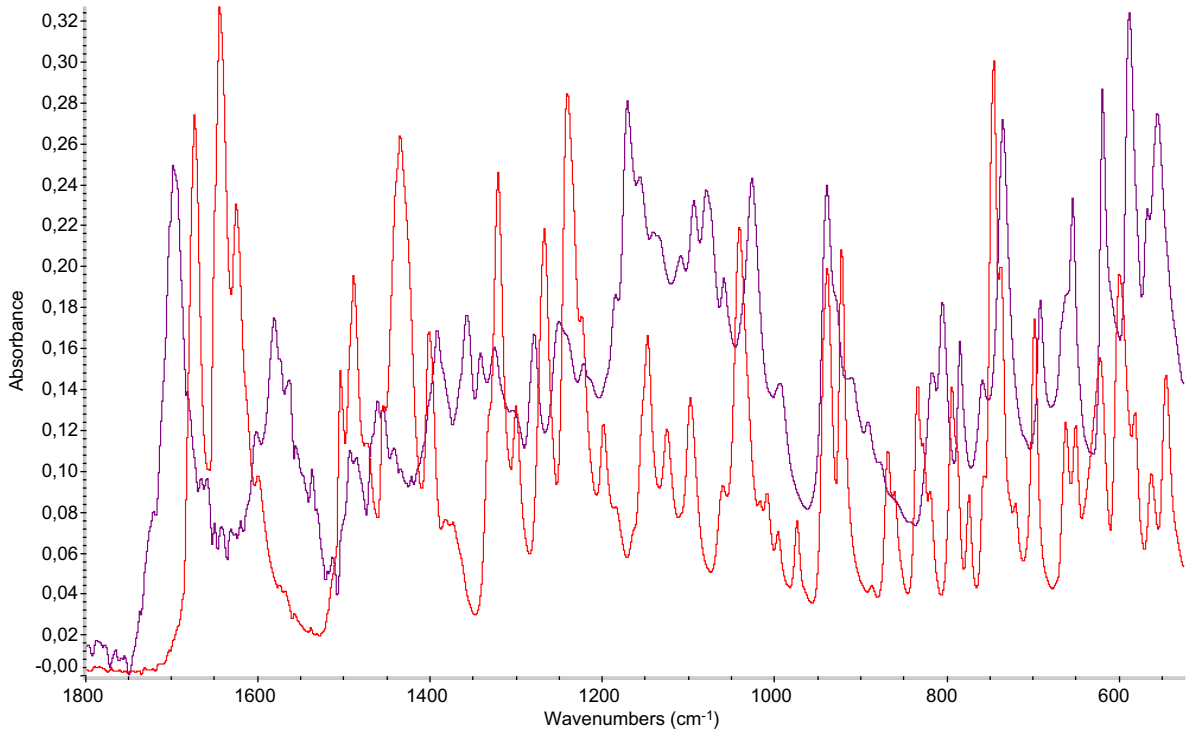

Figure 1 Representative spectra of SLD (in purple) and TAD (in red).

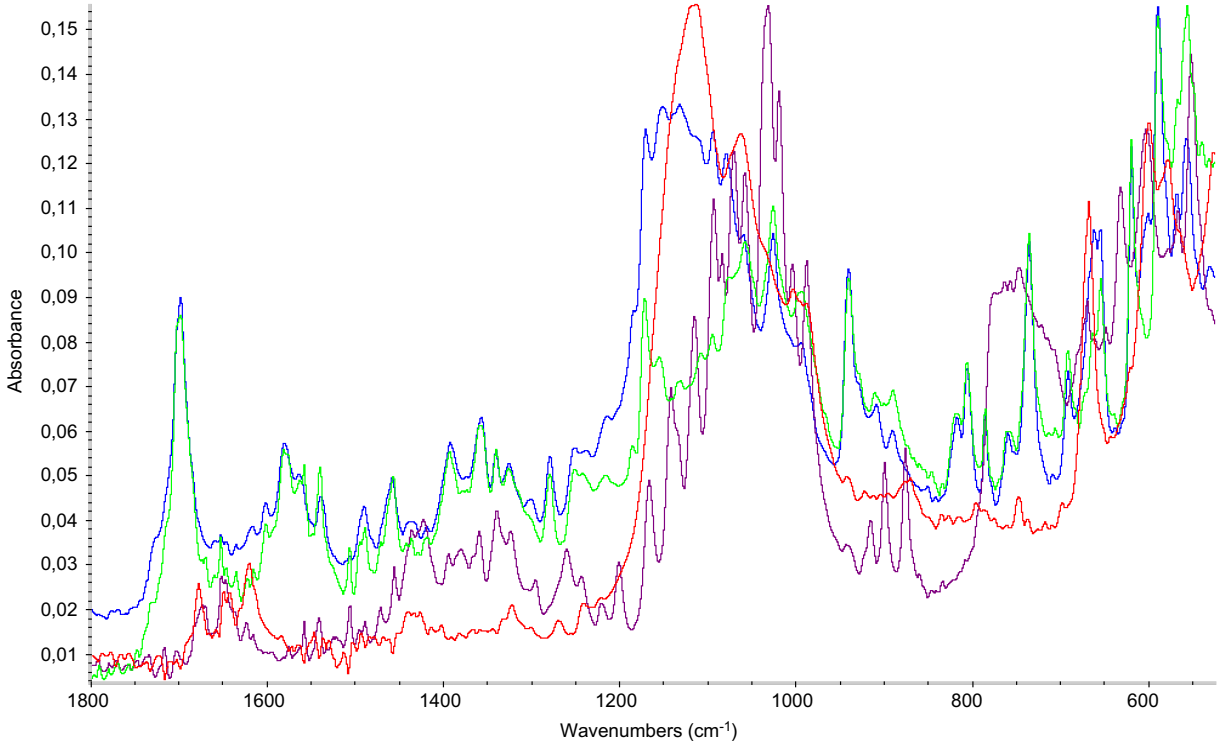

Figure 2 Representative spectra of Genuine Viagra (in green), Counterfeit Viagra (in blue), Genuine Cialis (in purple), and Counterfeit Cialis (in red).

2.3.1. Supervised Techniques

The *k*-Nearest Neighbor (KNN) technique categorizes a new sample in authentic or counterfeit classes by measuring the Euclidean distances between the new sample and the *k*-Nearest Neighbor, representing existing samples. The class of each of the *k* neighbors is known, authentic or counterfeit. A new

sample is classified as authentic if the majority of its k-Nearest Neighbors belong to the authentic class. The number of neighbors, k, can be defined by maximizing the classification accuracy in the training set. Further details in KNN are presented in Duda et al.[15] and Been et al..[13]

With similar purposes, the Support Vector Machine (SVM) is a classification tool that constructs a hyper plane to separate authentic and unauthentic samples by maximizing the distance between the two closest observations, one of authentic class and one of unauthentic class; see Cristianini and Shawe-Taylor.[16] In order to find a more precise separating hyper plane, original variables can be transformed by mathematical functions named kernel functions; such transformation reallocates the observations in a higher dimensional space that allows finding a separating hyper plane. Polynomial, Radial basis and Sigmoid are among the most used kernel functions, as claimed by Abe[17] and Rakotomamonjy.[18]

The third method tested is the Probabilistic Neural Network (PNN), a classification technique that takes into account the influence of all the existing observations to categorize a new observation into authentic or unauthentic class. It calculates the Euclidean distance between a new observation and each of the existing observations.[15] These distances are transformed by means of a standard exponential function, which scales the similarity between the new observation and each of the existing observations; such scaling is weighted by a sigma parameter. PNN then sums up the transformed values related to observations belonging to the authentic class separately from those coming from the unauthentic class. The new observation is assigned to the class with the highest summation. Further details can be found in Spetch.[19]

The last ST tested is the well acknowledged Linear Discriminant Analysis (LDA). LDA constructs a linear combination of variables that enables the classification of observations in two or more classes.[20,21] In LDA, the dependent variable is a categorical variable which identifies the class of each observation, and the coefficients of the discriminant function are defined in a way that the variance between the groups is maximized; see Abdi[22] for details on LDA.

2.3.2. Unsupervised Techniques

Data clustering is a widely known multivariate analysis technique that inserts observations (samples) into classes (clusters) so that observations in the same cluster are as similar as possible, and items in different clusters are as dissimilar as possible.[23,24] Clustering algorithms typically belong to two approaches: nonhierarchical and hierarchical methods. The k-Means clustering algorithm, one of the most hailed non-hierarchical methods,[25] inserts each observation into the cluster with the nearest centroid. The method aims at minimizing the sum of the Euclidean distances between the observations and the nearest centroid.[26] The number of clusters k is user-defined.

An alternative clustering technique is the Fuzzy C-Means (FCM), in which each observation has a degree belonging to clusters rather than belonging entirely to a single cluster. For that matter, FCM computes a metrics similar to a "membership grade" that measures how much each observation belongs to a cluster. That grade is inversely related to the distance from a specific observation to the cluster centroids around that observation.[27] Formally, one observation is inserted into the cluster that presents the higher probability of owing that observation, i.e., the cluster with the highest "membership grade". Nock and Nielsen[28] compared the performance of different FCMs; additional details on FCM are available in Ahmed et al.[27]

The quality of the clustering procedure can be assessed by the Silhouette Index (SI), which measures how similar an observation is with respect to observations in its own cluster, compared with observations in other clusters.[24,29] SI is estimated as in Eq. (1), where $a(j)$ is the average distance from the j-th observation to all others in its cluster, and $b(j)$ is the average distance from the j-th observation to all oth ers assigned to the nearest neighboring cluster.

$$SI_j = \frac{b(j) - a(j)}{\max\{b(j), a(j)\}} \quad (1)$$

Each clustered observation is associated with a SI value that ranges from +1 to −1; the closer to +1 the more distant the observation is to observations in neighboring clusters, meaning a proper cluster; values of SI close to 0 indicate observations that do not clearly belong to a specific cluster; SI values close to −1 refer to observations that were improperly inserted into the final cluster. Kaufman and Rousseeuw[24] state that the global quality of a clustering procedure can be assessed by estimating the average SI over all clustered observations.

Although not tailored to grouping purposes, Principal Components Analysis (PCA) belongs to the UT class. PCA is a multivariate technique that constructs A independent linear combinations of the original variables x. Consider data from an analytical technique consisting of N samples described by J variables (wave numbers); the linear combinations of the x variables are represented by $t_{ia} = w_{1a}x_{i1} + w_{2a}x_{i2} + \ldots + w_{Ja}x_{iJ}$. These combinations, also known as principal components, explain most of the variability in the original data based on a small number of components; i.e., $A < J$. The weight associated with variable j, w_{ja}, is defined based on the maximization of the variance between the components. In these propositions, the t variables, also called scores, replace the original x variables in the supervised and unsupervised techniques in order to improve grouping accuracy. The number of components to be retained is defined based on the amount of explained variance, as in Rencher.[30] For details about PCA, see Anzanello et al.[31]

3. Results and discussion

The supervised (KNN, LDA, PNN and SVM) and unsupervised techniques (k-Means CA and FCM) were applied to 28 samples of authentic Cialis, 25 samples of authentic Viagra, and 104 counterfeit samples sent to the PF (Porto Alegre, Brazil) for forensic analysis. The techniques were evaluated in terms of their ability to accurately insert the samples into two classes: authentic and unauthentic, i.e., the ST relied on the class label to train the model, while the UT was supposed to distinguish the two groups based on the data structure itself.

The six multivariate techniques were applied to the original data from FTIR (denoted as variables x), and to the scores t yielded by a PCA analysis run on the original data. Thus, the intention was to compare the performance of the techniques also on PCA scores, since that is a typical procedure

in forensic analysis.[7] In addition, the original variables x and scores t were transformed using two levels of a kernel polynomial transformation, x^3 and $x^{1/3}$ (similarly for t^3 and $t^{1/3}$), in order to evaluate whether data transformation improves the performance of the techniques. Kernel methods promote a transformation on the data, remapping the samples into a high-dimensional variable space. Such procedure usually reveals new structure in the data, yielding better classification and clustering results.[32]

The performance of the ST (KNN, LDA, PNN and SVM) was evaluated through the classification accuracy of samples on a testing set. For that matter, the original data-set was divided into training and testing sets according to a 75%/25% proportion.[33] The clustering performance was evaluated by rescaling the Silhouette Index (SI): since the original SI is comprised in the (−1 to 1) interval, it was adjusted to a percent scale (0 to 1). That enabled an unbiased comparison between the supervised and unsupervised techniques. Finally, a cross-validation procedure[15] recommended using $k = 3$ for the KNN, and sigma = 0.001 for the PNN. As for the PCA, 2 principal components were retained based on a scree graph.[30]

Table 1 depicts the performance of the six techniques applied to the Viagra data. The higher the percent, the better the classification or clustering results. The ST presented a satisfactory average accuracy (0.9308) when applied to the x variables, but that performance decreased significantly when the PCA scores replaced the original variables (average accuracy = 0.5628). A comparison between the four ST shows that the SVM slightly outperforms KNN and PNN; the LDA yields a significantly lower accuracy when compared with the other techniques. There is no clear tendency due to the kernel transformation on such results.

As for the unsupervised techniques, a lower average accuracy (0.8731) was observed when compared with the Supervised Techniques (0.9308) applied to the x variables (see Table 1). Both k-Means and FCM presented a similar performance. Conversely to the ST, the clustering methods presented a higher average accuracy (0.9279) when applied to the PCA scores than to the x variables (0.8731). That agrees with many forensic studies that claim that analyses on PCA scores yield more conclusive results compared with the information provided by the x variables.[7,34,35] It is also noteworthy that the cubic kernel transformation (x^3 and t^3) tends to increase the classification/clustering performance.

Similar analysis was applied to the Cialis data; results are presented in Table 2. Cialis data corroborated the supervised techniques as a better choice when multivariate techniques are to be applied on x variables: the average classification accuracy for the ST is 0.9558, while the UT yielded average 0.8214. The kernel transformations do not favor the ST, decreasing the average accuracy from 0.9558 to 0.7198. Once again, the LDA was outperformed by the other ST.

Table 1 Classification and clustering performance for the Viagra data.

Kernal	Supervised				Unsupervised	
	KNN	LDA	PNN	SVM	k-Means	FCM
$x^{1/3}$	0.9690	0.7854	0.9665	0.9995	0.8573	0.8573
x	0.9644	0.7817	0.9489	0.9998	0.8573	0.8573
x^3	0.9591	0.8439	0.9750	0.9769	0.9098	0.8995
$t^{1/3}$	0.5300	0.5475	0.6093	0.5922	0.9228	0.9216
t	0.5303	0.5555	0.5889	0.6270	0.9228	0.9216
t^3	0.5526	0.5218	0.6063	0.4925	0.9405	0.9381

Table 2 Classification and clustering performance for the Cialis data.

Kernal	Supervised				Unsupervised	
	KNN	LDA	PNN	SVM	k-Means	FCM
$x^{1/3}$	0.9992	0.8620	0.9885	0.9995	0.7608	0.7608
x	0.9995	0.8659	0.9493	0.9998	0.7608	0.7608
x^3	0.9995	0.8334	0.9964	0.9769	0.9435	0.9417
$t^{1/3}$	0.8313	0.4881	0.7619	0.8722	0.8688	0.8685
t	0.8495	0.4976	0.7370	0.8632	0.8688	0.8685
t^3	0.7369	0.5334	0.7204	0.7458	0.9689	0.9689

As verified for the Viagra data, the UT on the PCA scores presented a higher average accuracy (0.9021) when compared with the x variables (0.8214). In addition, the cubic kernel transformation (x^3 and t^3) substantially increased the clustering performance. Finally, both k-Means and FCM presented a similar performance.

Figures 3 and 4 display the SI graphs for the worst and best clustering results for the Cialis, respectively. There is a remarkable improvement on the clustering quality when using the cubic kernel aligned with the PCA scores: the number of misclassified samples is reduced in Figure 4, and a substantial number of samples yields SIs close to 3, i.e., denoting a proper clustering procedure.

Based on the aforementioned results, it seems reasonable to recommend three of the ST (KNN, PNN and SVM) to scenarios where the data are described by the original x variables; i.e., scenarios where PCA is not suitable for analysis. On the other hand, k-Means and FCM clustering techniques are favored by PCA scores, which significantly increased grouping performance. There is no clear distinction between the two UT tested, suggesting that both can be used without a major loss. The cubic kernel transformation also showed to be a valuable resource for unsupervised analysis. Finally, the computational time required by all the tested techniques is very small, and the codes are usually available in most statistical packages.

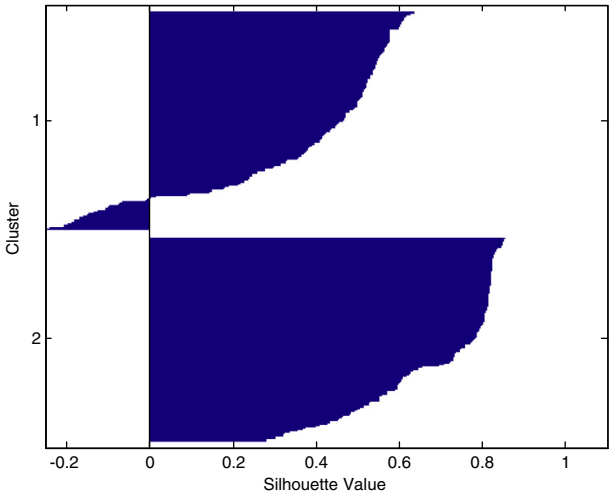

Figure 3 SI graph on $x^{1/3}$.

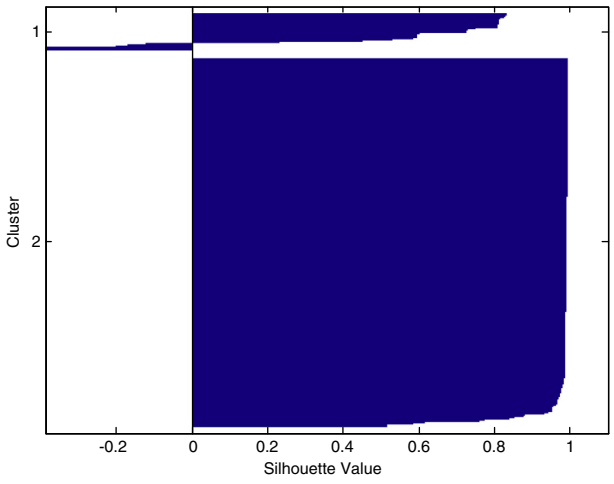

Figure 4 SI graph on t^3.

4. Conclusion

This paper compared the performance of two groups of multivariate techniques frequently used for analyzing sample properties and inserting samples into authentic and unauthentic categories: supervised and unsupervised techniques. For that matter, four STs (KNN, LDA, PNN and SVM) and two UTs (*k*-Means and FCM) were tested on FTIR from authentic and unauthentic Viagra and Cialis samples. The original data were also transformed by PCA and kernel functions aimed at trans-forming the data and improving the grouping performance.

STs proved to be a more reasonable choice when the analysis was conducted on the original data, i.e., *x* variables. KNN, PNN and SVM presented a better performance than LDA in both datasets, but neither PCA nor kernel transformations yielded better grouping results when integrated to a ST. As for the UT, *k*-Means and FCM performed similarly. It is noteworthy that the clustering improvement was yielded by the cubic kernel transformation, suggesting that remapping techniques unveil implicit patterns on data from analytical techniques.M.J. Anzanello et al.

Future studies include the development of variable selection approaches for both ST and UT, since better grouping results may derive from using a subset of more informative variables. Other kernel transformations, such as the Sigmoid and Gaussian, are also to be tested. The techniques will also be applied on data from other analytical techniques.

Funding

None.

Conflict of interest

None declared.

Ethical approval

Necessary ethical approval was obtained from the institute ethics committee.

References

1. Ortiz RS, Mariotti KC, Schwab NV, Sabin GP, Rocha WFC, Castro EVR, et al. Fingerprinting of Sildenafil Citrate and Tadalafil tablets in pharmaceutical formulations via X-ray fluorescence spectrometry XRF. *J Pharmaceut Biomed* 2012;**58**:7–11.
2. Holzgrabe U, Malet-Martino M. Analytical challenges in drug counterfeiting and falsification-the NMR approach. *J Pharmaceut Biomed* 2011;**55**:679–87.
3. Ortiz RS, Mariotti KC, Limberger RP, Mayorga P. Physical profile of counterfeit tablets Viagra and Cialis. *Braz J Pharm Sci* 2012;**48**:1–9.
4. Ortiz RS, Mariotti K, Roma~o W, Eberlin MN, Limberger RP, Mayorga P. Chemical fingerprinting of counterfeits of Viagra and Cialis tablets and analogues via electrospray ionization mass spectrometry. *Am J Anal Chem* 2011;**2**:919–28.
5. Planinsek O, Planinsek D, Zega A, Breznik M, Srcic S. Surface analysis of powder binary mixtures with ATR FTIR spectroscopy. *Int J Pharm* 2006;**319**:13–9.
6. López-Sánchez M, Domínguez-Vidal A, Ayora-Cañada MJ, Molina-Dıaz A. Assessment of dentifrice adulteration with diethylene glycol by means of ATR-FTIR spectroscopy and chemometrics. *Anal Chim Acta* 2008;**620**:113–9.
7. Ortiz RS, Mariotti K, Fank B, Limberger R, Anzanello MJ, Mayorga P. Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: can the same pharmaceutical powder mixture be used to falsify two medicines? *Forensic Sci Int* 2013;**226**:282–9.
8. Thanasoulias NC, Parisis NA, Evmiridis NP. Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. *Forensic Sci Int* 2003;**138**(1–3):75–84.
9. Brewer LN, Ohlhausen JA, Kotula PG, Michael JR. Forensic analysis of bioagents by X-ray and TOF-SIMS hyperspectral imaging. *Forensic Sci Int* 2008;**179**(2–3):98–106.
10. Campbell GP, Curran JM, Miskelly GM, Coulson S, Yaxley GM, Grunsky EC, et al. Compositional data analysis for elemental data in forensic science. *Forensic Sci Int* 2009;**188**(1–3):81–90.
11. Den Hartog BK, Elling JW. Clustering for forensic mitotype quality analysis. *Forensic Sci Int, Genetics Supplement Series* 2009;**2**(1):317–9.
12. Broséus J, Anglada F, Esseiva P. The differentiation of fibre- and drug type Cannabis seedlings by gas chromatography/mass spectrometry and chemometric tools. *Forensic Sci Int* 2010;**200**(1–3):87–92.
13. Been F, Roggo Y, Degardin K, Esseiva P, Margot P. Profiling of counterfeit medicines by vibrational spectroscopy. *Forensic Sci Int* 2011;**211**(1–3):83–100.
14. Sikirzhytski V, Sikirzhytskaya A, Lednev IK. Advanced statistical analysis of Raman spectroscopic data for the identification of body fluid traces: semen and blood mixtures. *Forensic Sci Int* 2012;**222**:259–65.
15. Duda R, Hart P, Stork D. *Pattern classification*. 2nd ed. New York: Wiley-Interscience; 2001.
16. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press; 2000.
17. Abe S. *Support Vector Machine for pattern classification*. London: Springer; 2005.
18. Rakotomamonjy A. Variable Selection using SVM-based criteria. *J Mach Learn Res* 2003;**3**:1357–70.
19. Specht DF. Probabilistic neural networks. *Neural networks* 1990;**3**:109–18.
20. Daéid N, Waddell R. The analytical and chemometric procedures used to profile illicit drug seizures. *Talanta* 2005;**67**:280–5.
21. Williams MR, Sigman ME, Lewis J, Pitan KM. Combined target factor analysis and Bayesian soft-classification of interference-

contaminated samples: Forensic Fire Debris Analysis. *Forensic Sci Int* 2012;**222**:373–86.
22. Abdi H. Discriminant correspondence analysis. In: Salkind NJ, editor. *Encyclopedia of measurement and statistic*. Thousand Oaks (CA): Sage; 2007. p. 270–5.
23. Jobson J. *Applied multivariate data analysis, V. II: categorical and multivariate methods*. New York: Springer-Verlag; 1992.
24. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*. New Jersey: Wiley Interscience; 2005.
25. Jain A, Dubes R. *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall; 1988.
26. Taboada H, Coit D. Data clustering of solutions for multiple objective system reliability optimization problems. *Qual Technol Quant M* 2007;**4**(2):191–210.
27. Ahmed MN, Yamany SM, Mohamed N, Farag AA, Moriarty T. A modified fuzzy C-means algorithm for bias field estima-tion and segmentation of MRI data. *IEEE T Med Imaging* 2002;**21**:193–9.
28. Nock R, Nielsen F. On Weighting Clustering'', IEEE T Pattern Anal 2006;**28** (8):1–13.
29. Anzanello M, Fogliatto F. Selecting the best clustering variables for grouping mass-customized products involving workers learning. *Int J Prod Econ* 2011;**130**:268–76.
30. Rencher A. *Methods of multivariate analysis*. New York: Wiley Interscience; 1995.
31. Anzanello M, Fogliatto F, Rossini K. Data mining-based method for identifying discriminant attributes in sensory profiling. *Food Qual Prefer* 2011;**22**:139–48.
32. Schölkopf B, Smola AJ. *Learning with Kernels*. Cambridge, MA: MIT Press; 2002.
33. Anzanello M, Albin S, Chaovalitwongse W. Selecting the best variables for classifying production batches into two quality levels. *Chemom Intell Lab Syst* 2009;**97**:111–7.
34. Muehlethaler C, Massonnet G, Esseiva P. The application of chemometrics on Infrared and Raman spectra as a tool for the forensic analysis of paints. *Forensic Sci Int* 2011;**209**:173–82.
35. Rajalahti T, Kvalheim OM. Multivariate data analysis in pharmaceutics: a tutorial review. *Int J Pharm* 2011;**417**:280–90.